MULTIVARIATE ANALYSIS APPLIED TO THE CALIFORNIA HEALTH INTERVIEW SURVEY

An Undergraduate Research Scholars Thesis

by

AARON ROSS

Submitted to the Undergraduate Research Scholars program at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by Research Advisor: Dr. Dennis Jansen

May 2019

Major: Economics

TABLE OF CONTENTS

ABSTRACT

Multivariate Analysis Applied to the California Health Interview Survey

Aaron Ross
Department of Economics
Texas A&M University

Research Advisor: Dr. Dennis Jansen
Department of Economics
Texas A&M University

Objective

Identify if principle components analysis and multiple correspondence analysis are suitable dimension reduction techniques for the California Health Interview Survey. Identify which health risk behaviors, mental health and demographic factors cluster utilizing k-medians clustering.

Background

Clustering and multivariate analysis techniques can be used to characterize populations and sub-populations of people by grouping them based on an individual's similarity to others. These exploratory techniques, while uniformly accepted within the scientific community as valid, are not as popular as other statistical methods and have not been utilized in certain scenarios where they could potentially be useful. The UCLA Center for Health Policy Research's annual California Health Interview Survey (CHIS) dataset is one such example where using these multivariate techniques could provide new insight. The survey contains information on thousands of randomly sampled Californians regarding health, income and demographics, among other factors. This research project attempts to determine if principle components analysis and

multiple correspondence analysis are suitable dimension reduction techniques when applied to the CHIS dataset and to quantify and qualify in greater detail the differences and similarities between the health characteristics of California residents.

Methods

This study used data from 21,055 individuals interviewed via telephone from the 2016 California Health Interview Survey, the largest state-wide health survey in the U.S. The statistical procedures principle components analysis and multiple correspondence analysis were conducted to assess their usefulness when applied to health survey data. Concurrently, Gower k-medians clustering was used to identify distinct groupings of California residents. I then performed a chi-squared test to determine which variables are the most statistically significant in forming these clusters.

Results

Principle components analysis reduced the initial 118 variables considered to 30, with the largest component only explaining 10.44% of the total variation in the data, suggesting that this technique is ill-suited to the CHIS. Multiple correspondence analysis, however, reduced the 88 categorical variables to 5 with the largest component accounting for 62.27% of the variation in the data. By applying Gower k-medians, I produced 3 distinct clusters of survey respondents and determined that access to specialized medical care is the most strongly clustered characteristic.

INTRODUCTION

Great scrutiny has been given to the causal relationship between certain demographic factors, such as income and obesity (Kim & Knesebeck, 2018). However, given the complex nature and multi-dimensionality of current public health datasets, prescribing policy based on classifications of people considering only a few factors can lead to inefficient and suboptimal outcomes. However, considering too many factors can make the interpretation of results difficult and result in overfitting. Multivariate clustering and dimension reduction techniques are two methods that are useful in attempting to solve these problems by algorithmically grouping observations based on their similarity to each other and selecting the most important features or combination of features, respectively.

Clustering has been used in conjunction with public health data with success in the past; for example, a previous study published in 2014 called "Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area" examined and distinguished 4 distinct categories of people who utilize the health care system in metropolitan Paris (Lefèvre, Rondet & Parizot, 2014). Similarly, a study in 2003 using the CHIS dataset used clustering to discover that certain health risk behaviors and resiliency factors are associated with each other; for example, they discovered a strong correlation between parental supervision and overeating (Mistry et al., 2009).

There has been a long known association between health risk behaviors, such as smoking, using drugs, being primarily sedentary and negative health conditions, such as cancer, HIV and heart disease (Hruby & Hu, 2015). These negative health conditions lead to a decline in quality of life for the individual and often impose a negative externality of their suffering and the cost of their care on their families and society (Megari, 2013). Research has shown that many risk behaviors tend to cluster, that there exist different distinct risk behavioral groups that share certain characteristics. Additionally, relatively recent literature reveals a similar relationship between risk behaviors and mental health (Busch et al., 2013). Identifying and understanding the make-up of these clusters can be a critical tool in understanding and recognizing individuals most at risk. Successful examples using clustering to optimize or evaluate policy include using geospatial and Medicare spending data in conjunction with hierarchal clustering to discover eight distinct "service-usage patterns" and to better identify the unique needs of different groups of high-cost patients.

The usefulness of dimension reduction techniques in analyzing datasets with a large number of variables has been recognized in the past. Specifically, within the context of contemporary biomedical research, where datasets can often contain more variables than observations, dimensionality reduction has proved critical in transforming datasets into a format where traditional statistical analysis is computationally feasible (Lee et al., 2016). While most often applied to genomic data, these methods have also revealed essential features and patterns in non-genomic health datasets such as the diagnosis of heart disease (Shilaskar & Ghatol, 2013). The usefulness of this technique is not only limited to the analysis of biological science data. The emergence of big data has made the use of dimension reduction techniques helpful in examining datasets with a large number of socioeconomic factors as well. The study “Some dimension reduction strategies for the analysis of survey data” shows how the dimension reduction technique PCA can be employed specifically to survey data (Weng & Young, 2017). Additionally, the reduction technique MCA has been used to construct indices representing individuals physical and mental health (Kohn, 2012).

Intuitively, these techniques, clustering and dimension reduction attempt to reduce the “noise” in data sets by reducing the number of variables deemed crucial and identifying groups within the data, thereby simplifying interpretation. In contrast to typical deterministic models and statistical analysis, these exploratory techniques often have less rigorous assumptions and are non-casual. These characteristics make these techniques attractive to health researchers analyzing high dimensional and complex datasets.

California is the most populous state in the U.S. and one of the most culturally and economically diverse. Although touting the 6th largest GDP in the world and a growing economy, California’s income distribution has become more and more stratified, noting rising incomes among the well-off but drop among those with mid-to-lowest levels of income. Additionally, the study “Behavioral Health Barometer: California, 2014” conducted by the Substance Abuse and Mental Health Services Administration found some public health threats were undertreated. More specifically, only 8% of the population with an alcohol dependence or abuse problem and 36.5% of residences with a mental illness had ever received some sort of treatment (Behavioral Health Barometer: California, 2014). These issues make the state a prime candidate for conducting detailed exploratory techniques to better understand the unique needs of their residents.

Aimed at better understanding these issues, the annual California Health Interview Survey (CHIS) collects extensive information for all age groups in California on variety of health issues and their demographics. With a uniquely large questionnaire, financial incentives and data from every county, the CHIS aims to be the most comprehensive dataset of its kind in the U.S. Additionally, its combination of both health and demographic data provides us with uniquely large sample and variable list, which can provide special insight into the wellbeing of the residents in the state.

This paper aims to assess the adequacy of the dimensionality techniques principle components analysis and multiple correspondence analysis and attempts to identify clusters of individuals pertaining to physical and mental health, and other socioeconomic factors in application to the CHIS data. In doing this I aim to achieve the following three main goals: to provide insight into the prevalence and strength of groupings of Californians, to compare the results of several grouping techniques and identify which procedures are the most appropriate, and to create an example for other research to be conducted using these types of multivariate analysis.

I hope that in employing principle components analysis, multiple correspondence analysis, and clustering I will be able to create a foundation for future analysis to be conducted.

SECTION I

METHODS

Survey

The California Health Interview Survey (CHIS) is a collaborative project between the University of California, Los Angeles Center for Health Policy Research, the California Department of Public Health, the Department of Health Care Services, and the Public Health Institute. It is the largest state-wide health survey in the nation and serves a critical role in providing policymakers, researchers, members of the media and others reliable information on the well-being of Californians.

According to the UCLA Center for Health Policy, the sample is created to meet the following two primary objectives of providing estimates for large- and medium-sized counties in the state, and for groups of the smallest counties (based on population size) and providing statewide estimates for California's overall population, its major racial and ethnic groups, as well as several racial and ethnic subgroups.

The actual survey is conducted via telephone employing a multi-stage sample design aiming to achieve parity in the number of completed interviews by landline and cellular phones. The response rate to them respectively were 6.8 percent and 8.4 percent. To incentivize participation in the survey, monetary compensation, ranging from $5-$40, was offered to respondents.

Beneath is a table of the methods by which the survey was administered as well as a list of the populations studied for the 2015 and 2016 survey.

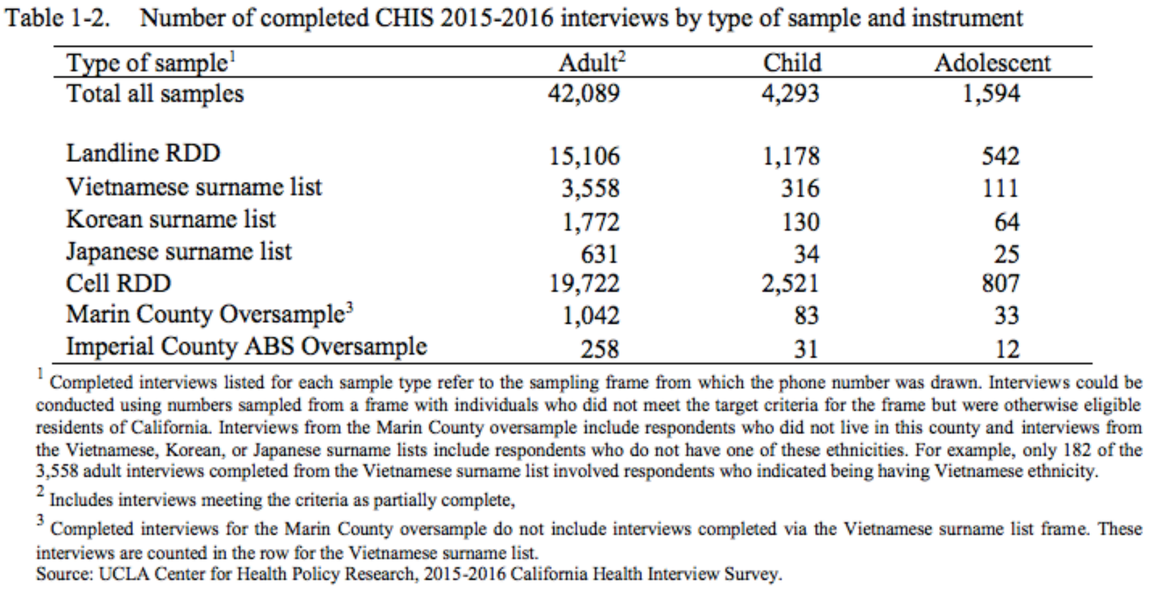


Table 1-2. Number of completed CHIS 2015-2016 interviews by type of sample and instrument

Type of sample[1]	Adult[2]	Child	Adolescent
Total all samples	42,089	4,293	1,594
Landline RDD	15,106	1,178	542
Vietnamese surname list	3,558	316	111
Korean surname list	1,772	130	64
Japanese surname list	631	34	25
Cell RDD	19,722	2,521	807
Marin County Oversample[3]	1,042	83	33
Imperial County ABS Oversample	258	31	12

[1] Completed interviews listed for each sample type refer to the sampling frame from which the phone number was drawn. Interviews could be conducted using numbers sampled from a frame with individuals who did not meet the target criteria for the frame but were otherwise eligible residents of California. Interviews from the Marin County oversample include respondents who did not live in this county and interviews from the Vietnamese, Korean, or Japanese surname lists include respondents who do not have one of these ethnicities. For example, only 182 of the 3,558 adult interviews completed from the Vietnamese surname list involved respondents who indicated being having Vietnamese ethnicity.

[2] Includes interviews meeting the criteria as partially complete,

[3] Completed interviews for the Marin County oversample do not include interviews completed via the Vietnamese surname list frame. These interviews are counted in the row for the Vietnamese surname list.

Source: UCLA Center for Health Policy Research, 2015-2016 California Health Interview Survey.

The CHIS's journal have published some studies using variations of these techniques, including one that quantifies the segmentation between the needs and costs of high-cost patients and low-cost patients in California (Davis, 2018), none have the scope of which my project proposes to do with this population and dataset.

Data

I used the adult subsample (N=21,055) of the 2016 CHIS dataset. In the survey, participants were asked a wide variety of questions regarding their health, ranging from their physical health, both overall and related to specific conditions (e.g. asthma, heart disease, etc.), their health-related behaviors, (e.g. dietary intake, cigarette usage, walking, etc.), and their mental health. Additionally, respondents were asked about their healthcare utilization and their demographic information. The answers to these questions give us a large mixture of quantitative and qualitative information for each individual. Given the nature of the questions a majority of the variables are categorical.

Missing or unanswered questions were imputed using various forms of interpolation, with most variables missing responses for less than 1% of the total respondents. All data is self-reported and thus might be potentially biased. Additionally, certain populations are typically more likely than others to over-represented in telephone surveys (i.e. elderly people, the sick). In conducting my analysis of the data, I did not adjust or weigh observations to account for the complex sample design. These limitations should be taken into consideration when interpreting the results.

Dimension Reduction

Dimension Reduction refers to the technique that reduces the number of variables or features in a dataset. Having fewer features is desirable as this can reduce the computation time of models and algorithms and result in more easily interpretable results. Dimension Reduction is done by either selecting a subset of the original variables (feature selection) or generating some smaller set of new variables from the old ones (feature extraction). Rather than simply getting rid of some of the variables being considered, feature extraction generates a whole new set of variables. Feature extraction also has the additional use of being able to reveal patterns and relationships between the old variables which make it a practical exploratory multivariate technique. We employ two dimension reduction techniques, principle components analysis and multiple correspondence analysis, in our attempt to characterize the data.

Principle components analysis (PCA) and multiple correspondence analysis (MCA) are feature extraction techniques which seek to transform a set of predictor variables into a set of linearly independent variables such that the first new linearly independent variable explains more than the second, the second more than the third, and so on (Weng & Young, 2017). PCA and MCA are similar in their goal of obtaining linear combinations of the data which explain the most information, using the fewest amount variables, about the data. However, they use different metrics to describe the information. PCA seeks generate a set of components that capture the most variance in the data, while MCA seeks to maximize the covariance of between the data. PCA can best interpreted when used on continuous data, but is sometimes useful when applied to mixed data as well, while MCA can only be applied to categorical data.

I employed PCA on the entire mixed dataset of the CHIS surveyed individuals 118 variables containing their socioeconomic and health information. After this, I conducted MCA on the subset of the 88 categorical variables. To analyze the usefulness and practicality of the techniques, I examined the proportion of variance explained by the new variables. I did this quantitatively, looking at the eigenvalues of the new variables generated from PCA and the principle inertia of each new variable generated by MCA. I then retained the variables that have eigenvalues greater than 1 and principle inertias greater than .2 (Costa et al., 2013). Qualitatively, I employed a scree plot, a common heuristic technique that plots the proportion of variance explained by each variable. I then used the scree test by visually assessing the “elbow” point of the scree plot and retained the factors to the left of this point.

Clustering

Clustering is a way of grouping together instances of similar data points into clusters. Typically, it is used within the context of data mining as an exploratory data analysis technique. In contrast to typical statistical methods, such as regression techniques, clustering does not rely on robust assumptions about the data. Additionally, it is algorithmic and non-deterministic, it simply groups data points based on their similarity instead of identifying a causal relationship between them. Because of these features, clustering is extraordinarily useful in identifying sub-groups within a heterogeneous population. I used clustering as my approach to examine and identify groupings of Californians. In clustering the CHIS data, I identified the characteristics which have the greatest "pull", or rather which sets of characteristics distinguish Californians the most. Depending on the type of clustering being conducted, different groups can emerge from the data. I employed the clustering techniques of k-medians to perform my clustering.

K-means is probably the most well-known and intuitively simple of all clustering algorithms. The method is an unsupervised learning procedure, meaning the purpose of employing it is simply to understand more about the underlying structure of the data rather than create a model that makes predictions. The specific goal of k-means being to partition the data into the most distinct set of a k number of clusters (Makles, 2012). To be performed, k-means requires a data set and a specified number (k) to indicate the number of clusters to be generated. The k-means procedure starts by assigning, either randomly or specified by the user, certain data points as centroids. Then the remaining observations are grouped according to the centroid they are closest to, until every data point is partitioned into one of the k clusters. After this, a new centroid is generated for each cluster by taking the average, or "mean" of the observations in that cluster. The observations are then re-clustered based on whatever new centroid they are closest to. This process repeats until the centroids stop changing, meaning that there are no changes in group membership of the observations (Armstrong et al., 2012). The end result gives each data point a cluster.

Typically, the metric of distance used to determine how observations are grouped is Euclidean:

$$distance\big((x, y), (a, b)\big) = \sqrt{(x-a)^2 + (y-b)^2}$$

However, this metric for measuring distance is only usable on scaled continuous data. Given that the majority of our data is categorical, I needed to use another metric to determine how similar the survey respondents are to one another. The Gower method of dissimilarity is way of dealing with this problem by using an appropriate method to measure distance for each data type (the Manhattan distance for continuous and ordinal data and the matching distance for binary data), and then scaling each distance to a value between 0 and 1.

I originally utilized a Euclidean K-means clustering algorithm; however, the nature of my data (having dozens of categorical variables) made these initial results unusable. Instead, I employed Gower K-medians to handle the set of mixed variables. I selected the last 4 observations of the data to be the initial "random" centroids to make my results reproducible.

After clustering the observations, I performed chi-squared tests on the new cluster membership variable and the original health and demographic variables. Then I used the chi-squared statistic associated with each test to determine statistical significance and strength of each of the CHIS variables in generating the clusters. Due to the size of the survey and the nature of clustering, most variables are expected to be statistically significant ($p<.01$) (Lin, Lucas, & Shmueli, 2013). I present the variables with the highest chi-squared values i.e. the variables that are the most significant. All statistical analyses were performed with Stata 15.1.

SECTION II

RESULTS

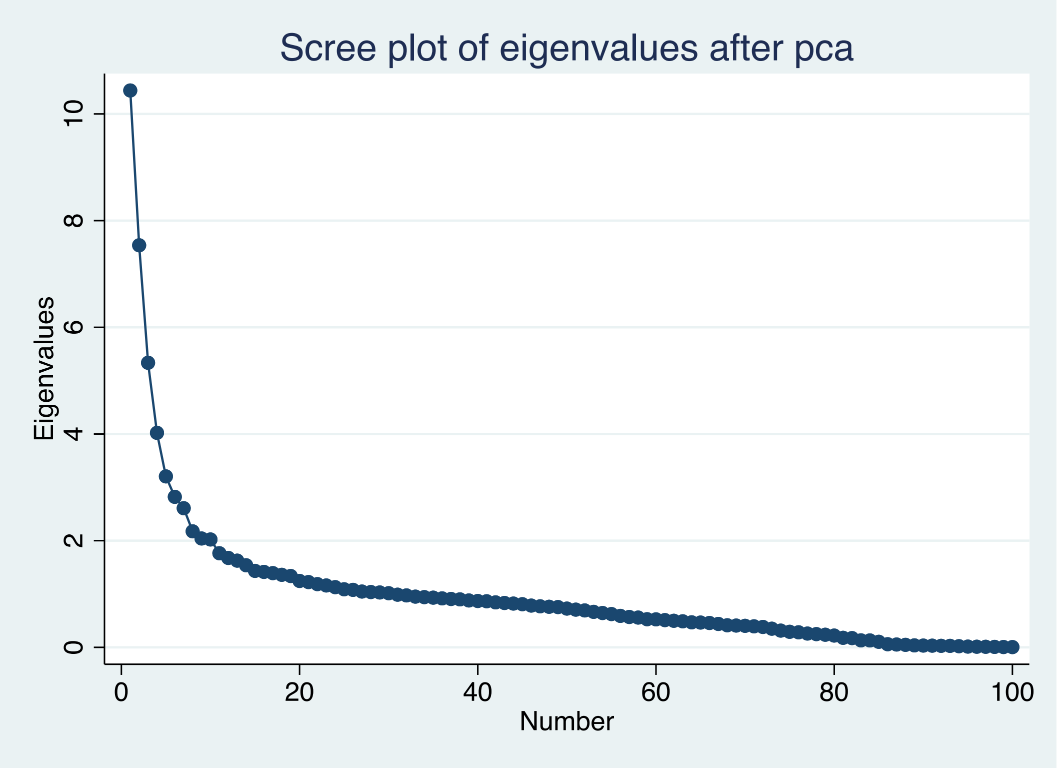
Principle Components Analysis

I first conducted PCA on the 2016 CHIS data set of 118 variables for the purpose of seeing if any dimension reduction was possible. Using independent orthogonal linear combinations of our variables, I was able to reduce the number variables from 118 to 30 utilizing the Kaiser's Criterion, which states that principle components (PCs) with eigenvectors greater than 1.0 should be retained.

```
Principal components/correlation                Number of obs    =     21,055
                                                Number of comp.  =          5
                                                Trace            =        100
    Rotation: (unrotated = principal)           Rho              =     0.3054

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      10.4403      2.90245             0.1044       0.1044
           Comp2 |      7.53782      2.20276             0.0754       0.1798
           Comp3 |      5.33507      1.31313             0.0534       0.2331
           Comp4 |      4.02194      .816595             0.0402       0.2734
           Comp5 |      3.20534      .383258             0.0321       0.3054
           Comp6 |      2.82208      .213291             0.0282       0.3336
           Comp7 |      2.60879      .433592             0.0261       0.3597
           Comp8 |       2.1752      .133745             0.0218       0.3815
```

Above are the first 8 PCs which explain 38% of the total variation between all of the variables. The proportion of variance explained by each PC drop dramatically with each new component constructed. The first two components alone make up half of the variation of the first eight components, while simultaneously only explaining 18% of the variation in the data. The first and most significant component, explaining 10% of the variation in the data, is positively related to overall health condition and negatively related to negative health-related behaviors.



The large number of PCs suggest that the PCA is not an effective tool in understanding the CHIS dataset, as the interpretation of many PCs is typically ineffective and less efficient than interpreting the original dataset. The health and demographic features are thus significantly different such that they cannot be reduced to a small number of variables by principle components analysis without significant loss of the proportion of variance explained.

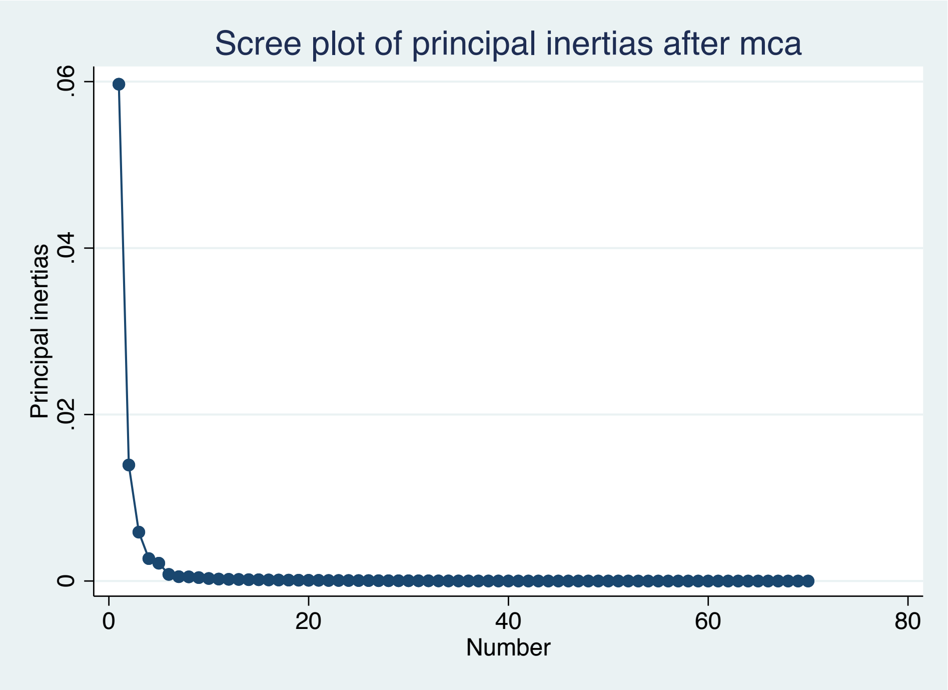
Multiple Correspondence Analysis

In addition to PCA, I employed multiple correspondence analysis (MCA) on our 88 categorical variables. MCA can be seen as an extension of correspondence analysis and a specific generalization of PCA where the variables are all categorical. Below are the first 8 components or dimensions generated from our original variables.

```
Multiple/Joint correspondence analysis         Number of obs     =     21,055
                                               Total inertia     =  .09586467
    Method: Burt/adjusted inertias             Number of axes    =          2

                 |   principal                cumul
      Dimension  |    inertia     percent    percent
    -------------+-----------------------------------
           dim 1 |    .0596909     62.27      62.27
           dim 2 |     .013943     14.54      76.81
           dim 3 |    .0058725      6.13      82.94
           dim 4 |    .0026957      2.81      85.75
           dim 5 |    .0021399      2.23      87.98
           dim 6 |    .0007956      0.83      88.81
           dim 7 |     .000521      0.54      89.35
           dim 8 |    .0005019      0.52      89.88
```

Above, the first and second dimensions contribute 62.27% and 14.54% respectively, and 76.81% cumulatively of the proportion of the total variation in survey data. In stark contrast to PCA, MCA does an excellent job of dimension reduction. Using the scree plot below and elbow heuristic we can see that the original 88 variables can be reduced to 5 components.



Gower K-Medians Cluster Analysis

I performed Gower k-medians cluster analysis on the mixed CHIS data. I chose k to be 3, based on a pseudo F-Statistic of 596. The algorithm partitioned the data into 3 clusters each containing the following number of observations:

```
----------------------
  clus_3 |      Freq.
----------+-----------
       1 |     11,112
       2 |      3,954
       3 |      5,989
----------------------
```

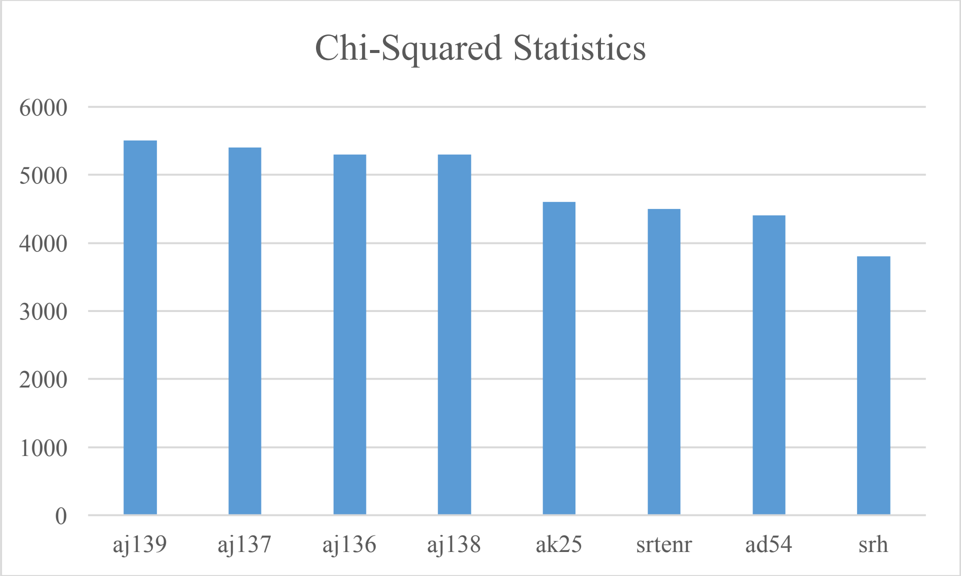
After partitioning the data into the 3 clusters, I conducted the chi-squared test for statistical significance. I found that the relationship between the clusters and every categorical variable to be statically significant (p<.01) with chi-squared statistics ranging from 17 to 8100. Below are the 10 variables ranked in order of significance, that have the largest chi-squared statistics:

Variable definitions

Variable	Variable Description	Variable Population	χ^2 statistic
aj139	If insurance was not accepted by specialist	Insured needing specialty care	5500
a137	Had trouble finding specialty doctor	Surveyees needing specialty care	5400
aj136	Needed to see medical specialist	All surveyees	5300
aj138	Not accepted as new patient by specialist	Surveyees needing specialty care	5300
ak25	Own or rent home	All surveyees	4600
srtenr	Self-reported household tenure	All surveyees	4500
ad54	Has difficulty working job	All surveyees under 65	4400
srh	Self-reported Latino or Hispanic	All surveyees	3800

Note: Variable Population indicates the Californians who qualify to answer the question. Adults who are not in the Variable Population do not qualify and are recorded as a separate category labeled inapplicable.

Below is a visual representation illustrating the size of each variable's chi-squared statistics:



The 4 most statistically significant variables (aj139, 1j137, aj136, aj138) all are in regard to an individual's need to receive or ability to receive specialized care. The two next most statistically significant variables (ak25, srtenr) indicate whether an individual owns, rents,or has some sort of other arrangement for their home. The 7th and 8th most significant variables (ad54, srh) indicate whether an individual has difficulty working at a job or if an individual is Latinx or Hispanic, respectively.

Below, I examined the prevalence of insured adults needing specialized care that is not covered by their insurance (aj139):

```
Proportion estimation: Cluster 1    Number of obs   =     11,112
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj139        |
INAPPLICABLE |   .4959503   .0047431      .4866555    .5052479
         YES |   .0357271   .0017608      .0324318    .0393438
          NO |   .4683225   .0047337      .4590557    .4776113
--------------------------------------------------------------
```

```
Proportion estimation: Cluster 2    Number of obs   =      3,954
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj139        |
INAPPLICABLE |   .2008093   .0063709      .1886091     .213591
         YES |   .0872534   .0044879      .0788464    .0964629
          NO |   .7119373   .0072019      .6976142    .7258485
--------------------------------------------------------------
```

```
Proportion estimation: Cluster 3    Number of obs   =      5,989
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj139        |
INAPPLICABLE |   .9215228   .0034749      .9144344    .9280701
         YES |   .0160294   .0016228      .0131399    .0195417
          NO |   .0624478   .0031266       .056592    .0688653
--------------------------------------------------------------
```

Examining each grouping reveals a large disparity between the clusters regarding the number of individuals who were insured and needed specialty care. Considering this variable, cluster 1 is characterized by 49% of its members either not needing specialty care or not having insurance. In addition, 92% of the members of Cluster 3 report either not needing specialty care or not having insurance. Cluster 2 by comparison reports only 20% of its members not having insurance or needing specialty care, meaning 80% *did* need specialty care. In addition, note how cluster 2 has a much higher prevalence of individuals who have insurance that *does not* cover their specialized medical needs, almost 9%.

Below we examine the third most important variable (aj136), which indicates if an individual sought a medical specialist in the past year:

```
Proportion estimation: Cluster 1    Number of obs   =     11,112
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj136        |
         YES |   .5087293   .0047425      .4994312    .5180214
          NO |   .4912707   .0047425      .4819786    .5005688
--------------------------------------------------------------
```

```
Proportion estimation: Cluster 2     Number of obs   =      3,954
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj136        |
         YES |   .8088012   .0062538      .7962392    .8207631
          NO |   .1911988   .0062538      .1792369    .2037608
--------------------------------------------------------------
```

```
Proportion estimation: Cluster 3     Number of obs   =      5,989
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
aj136        |
         YES |   .0946736    .003783      .0875131    .1023542
          NO |   .9053264    .003783      .8976458    .9124869
--------------------------------------------------------------
```

Individuals in cluster 1 exhibited an almost equal likelihood of either needing or not needing specialized medical care in the past year. In stark contrast to this, 80% of individuals in cluster 2 needed to see a medical specialist, but only 9.4% of individuals in cluster 3 needed to. These results support the proposition that access to specialized medical care strongly clusters among Californians.

To further characterize the groups, we examine the variable indicating the general health quality of an individual (ab1), which was the 9th most significant variable with a chi-squared statistic of 3700:

```
Proportion estimation: Cluster 1    Number of obs   =     11,112
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
ab1          |
   EXCELLENT |   .2098632    .003863      .2023914    .2175357
   VERY GOOD |   .3673506   .0045733      .3584329      .37636
        GOOD |    .299496   .0043451      .2910487     .308082
        FAIR |   .0989021    .002832      .0934879    .1045937
        POOR |    .024388   .0014633      .0216784    .0274269
--------------------------------------------------------------


Proportion estimation: Cluster 2    Number of obs   =      3,954
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
ab1          |
   EXCELLENT |    .041477   .0031709      .0356871    .0481593
   VERY GOOD |   .1474456   .0056384      .1367302    .1588463
        GOOD |   .2865453   .0071905      .2726577    .3008476
        FAIR |   .3239757   .0074425      .3095586    .3387349
        POOR |   .2005564   .0063679      .1883623    .2133324
--------------------------------------------------------------


Proportion estimation: Cluster 3    Number of obs   =      5,989
--------------------------------------------------------------
             |                                   Logit
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
ab1          |
   EXCELLENT |   .1482718    .004592      .1394936    .1575013
   VERY GOOD |    .239105   .0055116      .2284682    .2500766
        GOOD |   .3650025    .006221       .352896    .3772822
        FAIR |     .21456   .0053046      .2043442     .225142
        POOR |   .0330606   .0023104      .0288192    .0379019
--------------------------------------------------------------
```

From the results above, it should be noted how asymmetrical the average quality of health is for each of the clusters. Cluster 1 contains mostly healthy individuals with only 12.2% of these respondents having “fair” or “poor” health. In cluster 2, however, 52.3% of individuals report having “fair” or “poor” health. Cluster 3 reports 24.7% of individuals having “fair” or “worse” health, with the “fair” individuals making up 21.4% of that total cluster. With respect to this variable, the distribution of individuals in cluster 1 and 3 are heavily skewed in opposite direction, while the dispersion of individuals in cluster 2 is relatively normal, though the cluster is slightly skewed in the same direction as cluster 1.

CONCLUSION

In this paper, I conducted the dimension reduction techniques of principle components analysis and multiple correspondence analysis on the 2016 California Health Interview Survey. Principle components was found to be insufficient in reducing the total number of variables while multiple correspondence analysis was able to reduce the 88 categorical variables to 5 and maintaining 87.98% of the total variation in the original data. In addition to testing dimension reduction techniques I conducted Gower k-medians cluster analysis. Partitioning the data into three groups, and then conducting chi-squared tests for significance on the formed groupings and the categorical variables, I found the most important variables in generating the 3 clusters pertained to access to and the need for specialized medical care.

WORKS CITED

Abdi, H. & Valentin, D. (2007). Multiple Correspondence Analysis. In N. Salkind (ed.), Encyclopedia ofMeasurement and Statistics (pp. 651–657). Thousand Oaks: Sage

California Health Interview Survey. CHIS 2016 Adult Survey. UCLA Center for Health Policy Research. Los Angeles, CA: October 2017

Armstrong, J. J., Zhu, M., Hirdes, J. P., & Stolee, P. (2012). K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Archives of Physical Medicine and Rehabilitation, 93(12), 2198-2205. doi:10.1016/j.apmr.2012.05.026

Busch, V., Van Stel, H. F., Schrijvers, A. J., & de Leeuw, J. R. (2013). Clustering of health-related behaviors, health outcomes and demographics in Dutch adolescents: a cross-sectional study. *BMC public health*, *13*, 1118. doi:10.1186/1471-2458-13-1118

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa, N. (2013). The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing. Journal of Aging Research, 2013, 1-12. doi:10.1155/2013/302163

Davis, Shen. (2018). Segmentation of High-Cost Adults in an Integrated Healthcare System Based on Empirical Clustering of Acute and Chronic Conditions. *Journal of General Internal Medicine*.

Friesen, Elizabeth, Seliske, Papadopoulos. (2016). Using Principal Component Analysis to Identify Priority Neighbourhoods for Health Services Delivery by Ranking Socioeconomic Status. *Online Journal of Public Health Informatics PMC*.

Lee, G., Romo Bucheli, D. E., & Madabhushi, A. (2016). Adaptive Dimensionality Reduction with Semi-Supervision (AdDReSS): Classifying Multi-Attribute Biomedical Data. PloS one, 11(7), e0159088. doi:10.1371/journal.pone.0159088

Lefèvre, Rondet, Parizot. (2014). Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area.

Hruby, A., & Hu, F. B. (2015). The Epidemiology of Obesity: A Big Picture. PharmacoEconomics, 33(7), 673-89.

Jennifer L Kohn. (2012). "What is Health? A Multiple Correspondence Health Index," Eastern Economic Journal, Palgrave Macmillan; Eastern Economic Association, vol. 38(2), pages 223-250.

Kim TJ, Knesebeck O. (2018). Income and obesity: what is the direction of the relationship? A systematic review and meta-analysis. *BMJ Open*

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research Commentary—Too Big to Fail: Large Samples and thep-Value Problem. Information Systems Research, 24(4), 906-917. doi:10.1287/isre.2013.0480

Makles, A. (2012). Stata Tip 110: How to Get the Optimal K-Means Cluster Solution. The Stata Journal: Promoting Communications on Statistics and Stata, 12(2), 347-351. doi:10.1177/1536867x1201200213

Megari K. (2013). Quality of Life in Chronic Disease Patients. Health psychology research, 1(3), e27. doi:10.4081/hpr.2013.e27

Mistry, Ritesh et al. (2009). Resilience and Patterns of Health Risk Behaviors in California Adolescents. *Preventive medicine*, *PMC*

Soder, O. (2008, May 29). How does k-means clustering work? Retrieved from http://www.fon.hum.uva.nl/praat/manual/k-means_clustering.html

Soranzo, Vargas, Pascual-Montano. (2014). A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877

Substance Abuse and Mental Health Services Administration. Behavioral Health Barometer: California, (2014). HHS Publication No. SMA–15–4895CA. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2015.

Swati Shilaskar and Ashok Ghatol. (2013). Article: Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease. International Journal of Computer Applications 61(5):1-8

Weng, J., & Young, D. S. (2017). Some dimension reduction strategies for the analysis of survey data. Journal of Big Data, 4(1). doi:10.1186/s40537-017-0103-6